Size bias, sampling, the waiting time paradox, and infinite divisibility: when is the increment independent?

Richard Arratia, Larry Goldstein

July 23, 2010

Abstract

With X^* denoting a random variable with the X-size bias distribution, what are all distributions for X such that it is possible to have $X^* = X + Y$, $Y \geq 0$, with X and Y *independent*? We give the answer, due to Steutel [17], and also discuss the relations of size biasing to the waiting time paradox, renewal theory, sampling, tightness and uniform integrability, compound Poisson distributions, infinite divisibility, and the lognormal distributions.

1 The Waiting Time Paradox

Here is the "waiting time paradox," paraphrased from Feller [9], volume II, section I.4: Buses arrive in accordance with a Poisson process, so that the interarrival times are given by independent random variables, having the exponential distribution $\mathbb{P}(X > s) = e^{-s}$ for $s > 0$, with mean $\mathbb{E}X = 1$. I now arrive at an arbitrary time t. What is the expectation $\mathbb{E}W_t$ of my waiting time W_t for the next bus? Two contradictory answers stand to reason: (a) The lack of memory of the exponential distribution, i.e. the property $\mathbb{P}(X > r + s | X > s) = \mathbb{P}(X > r)$, implies that $\mathbb{E}W_t$ should not be sensitive to the choice t, so that $\mathbb{E}W_t = \mathbb{E}W_0 = 1$. (b) The time of my arrival is "chosen at random" in the interval between two consecutive buses, and for reasons of symmetry $\mathbb{E}W_t = 1/2$.

The resolution of this paradox requires an understanding of size biasing. We will first present some simpler examples of size biasing, before returning to the waiting time paradox and its resolution.

Size biasing occurs in many unexpected contexts, such as statistical estimation, renewal theory, infinite divisibility of distributions, and number theory. The key relation is that to size bias a sum with independent summands, one needs only size bias a single summand, chosen at random.

2 Size Biasing in Sampling

We asked students who ate lunch in the cafeteria "How many people, including yourself, sat at your table?" Twenty percent said they ate alone, thirty percent said they ate with one other person, thirty percent said they ate at a table of three, and the remaining twenty percent said they ate at a table of four. From this information, would it be correct to conclude that twenty percent of the tables had only one person, thirty percent had two people, thirty percent had three people, and twenty percent had four people?

Certainly not! The easiest way to think about this situation is to imagine 100 students went to lunch, and we interviewed them all. Thus, twenty students ate alone, using 20 tables, thirty students ate in pairs, using 15 tables, thirty students ate in trios, using 10 tables, and twenty students ate in groups of four, using 5 tables. So there were $20+15+10+5 = 50$ occupied tables, of which forty percent had only one person, thirty percent had two people, twenty percent had three people, and ten percent had four people.

A probabilistic view of this example begins by considering the experiment where an occupied table is selected at random and the number of people, X, at that table is recorded. From the analysis so far, we see that since 20 of the 50 occupied tables had only a single individual, $\mathbb{P}(X = 1) = .4$, and so forth. A different experiment, one related to but not to be confused with the first, would be to select a person at random, and record the total number X^* at the table where this individual had lunch. Our story *began* with the information $\mathbb{P}(X^* = 1) = .2$, $\mathbb{P}(X^* = 2) = .3$, and so forth, and the distributions of the random variables X and X^* are given side by side in the following table:

k	$\mathbb{P}(X=k)$	$\mathbb{P}(X^*=k)$
1	.4	.2
2	.3	.3
3	.2	.3
4	.1	.2
	1.0	1.0

The distributions of the random variables X and X^* are related; for X each table has the same chance to be selected, but for X^* the chance to select a table is proportional to the number of people who sat there. Thus $\mathbb{P}(X^* = k)$ is proportional to $k \times \mathbb{P}(X = k)$; expressing the proportionality with a constant c we have $\mathbb{P}(X^* = k) = c \times \mathbb{P}(X = k)$. Since $1 = \sum_k \mathbb{P}(X^* = k) = c \sum_k k\mathbb{P}(X = k) = c\mathbb{E}X$, we have $c = 1/\mathbb{E}X$ and

$$\mathbb{P}(X^* = k) = \frac{k\mathbb{P}(X = k)}{\mathbb{E}X}; \quad k = 0, 1, 2, \ldots. \tag{1}$$

Since the distribution of X^* is weighted by the value, or size, of X, we say that X^* has the X *size biased distribution*.

In many statistical sampling situations, like the one above, care must be taken so that one does not inadvertently sample from the size biased distribution in place of the one intended. For instance, suppose we wanted to have information on how many voice telephone lines are connected at residential addresses. Calling residential telephone numbers by random digit dialing and asking how many telephone lines are connected at the locations which respond is an instance where one would be observing the size biased distribution instead of the one desired. It's three times more likely for a residence with three lines to be called than a residence with only one. And the size bias distribution never has any mass at zero, so no one answers the phone and tells a surveyor that there are no lines at the address just reached! But the same bias exists more subtly in other types of sampling more akin to the one above: what if we were to ask people at random how many brothers and sisters they have, or how many fellow passengers just arrived with them on their flight from New York?

3 Size Bias in General

The examples in Section 2 involved nonnegative integer valued random variables. In general, a random variable X can be size biased if and only if it is nonnegative, with finite and positive mean, i.e. $1 = \mathbb{P}(X \geq 0)$ and $0 < \mathbb{E}X < \infty$. We will henceforth assume that X is nonnegative, with $a := \mathbb{E}X \in (0, \infty)$. For such X, we say X^* has the X size biased distribution if and only for all bounded continuous functions g,

$$\mathbb{E}g(X^*) = \frac{1}{a}\ \mathbb{E}(Xg(X)). \tag{2}$$

It is easy to see that, as a condition on distributions, (2) is equivalent to

$$dF_{X^*}(x) = \frac{x\, dF(x)}{a}.$$

In particular, when X is discrete with probability mass function f, or when X is continuous with density f, the formula

$$f(x) = \frac{xf(x)}{a}, \tag{3}$$

applies; (1) is a special case of the former.

If (2) holds for all bounded continuous g, then by monotone convergence it also holds for any function g such that $\mathbb{E}|Xg(X)| < \infty$. In particular, taking $g(x) = x^n$, we have

$$\mathbb{E}(X^*)^n = \mathbb{E}X^{n+1}/\mathbb{E}X \tag{4}$$

whenever $\mathbb{E}|X^{n+1}| < \infty$. Apart from the extra scaling by $1/\mathbb{E}X$, (4) says that the sequence of moments of X^* is the sequence of moments of X, but shifted by one. One way to recognize size biasing is through the "shift of the moment sequence;" we give an example in Section 15.

In this paper, we ask and solve the following problem: what are all possible distributions for $X \geq 0$ with $0 < \mathbb{E}X < \infty$, such that there exists a coupling in which

$$X^* = X + Y, \quad Y \geq 0, \qquad \text{and } X, Y \text{ are independent.} \tag{5}$$

Resolving this question on independence leads us to the infinite divisible and compound Poisson distributions. These concepts by themselves can be quite technical, but in our size biasing context they are relatively easy. We also present some background information on size biasing, in particular how it arises in applications including statistics. The answer to (5) comes from Steutel 1973 [17]; see section 10 for more of the history.

A beautiful treatment of size biasing for branching processes is [14] by Lyons, Pemantle, and Peres. Size biasing has a connection with Stein's method for obtaining error bounds when approximating the distributions of sums by the Normal, ([4] Baldi, P. Rinott, Y. 1989, [5] Baldi, P. Rinott, Y. and Stein C., 1989, and [10] Goldstein and Rinott, 1996), and the Poisson ([6], Barbour, Holst, and Janson, 1992).

To more fully explain the term "increment" in the title, letting $g(x) = \mathbb{1}(x > t)$ in (2) for some fixed t, we find that

$$\mathbb{P}(X^* > t) = \frac{1}{a}\ \mathbb{E}(X\mathbb{1}(X > t))\ \geq\ \frac{1}{a}\ \mathbb{E}X\ \mathbb{E}\mathbb{1}(X > t) = \mathbb{P}(X > t).$$

The inequality above is the special case $f(x) = x$, $g(x) = \mathbb{1}(x > t)$ of Chebyschev's correlation inequality: $\mathbb{E}(f(X)g(X)) \geq \mathbb{E}f(X)\ \mathbb{E}g(X)$ for any random variable and any two increasing functions f, g. The condition $\mathbb{P}(X^* > t) \geq \mathbb{P}(X > t)$ for all t is described as "X^* lies above X in distribution," and implies that there exist couplings of X^* and X in which always $X^* \geq X$. Writing Y for the difference, we have

$$X^* = X + Y, \quad Y \geq 0. \tag{6}$$

The simplest coupling satisfying (6) is based on the "quantile transformation," constructing each of X and X^* from the same uniform random variable U on (0,1). Explicitly, with cumulative distribution function F defined by $F(t) := \mathbb{P}(X \leq t)$, and its "inverse" defined by $F^{-1}(u) := \sup\{t : \ F(t) \leq u\}$, the coupling given by $X = F^{-1}(U), X^* = (F^*)^{-1}(U)$ satisfies (6).

In general (6) determines neither the joint distribution of X and Y, nor the marginal distribution of Y, nor whether or not X and Y are independent. It is a further restriction on the distribution of X to require that (6) be achievable with X, Y *independent.*

When $Z \sim Po(\lambda)$, i.e. Z is Poisson with $\mathbb{P}(Z = k) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, 2, \ldots$, we have $Z^* \overset{\mathrm{d}}{=} Z + 1$, where the notation $\overset{\mathrm{d}}{=}$ denotes equality in distribution. The reader can check

$$Z^* \overset{\mathrm{d}}{=} Z + 1 \tag{7}$$

directly using (2); a conceptual derivation is given in Example 1) in Section 16.1. Scaling by a factor $y > 0$ in general means to replace X

by yX, and it follows easily from (2) that

$$(yX)^* = y(X^*). \tag{8}$$

For our case, multiplying (7) by $y > 0$ yields the implication, for Poisson Z,

$$\text{if} \quad X = yZ, \quad \text{then} \quad X^* = X + y. \tag{9}$$

Hence, for each $\lambda > 0$ and $y > 0$, (9) gives an example where (5) is satisfied with Y a constant random variable, which is independent of *every* random variable. In a very concrete sense, all solutions of (5) can be built up from these examples, but to accomplish that we must first review how to size bias sums of independent random variables.

4 How to size bias a sum of independent random variables

Consider a sum $X = X_1 + \cdots + X_n$, with independent non-negative summands X_i, and suppose that $\mathbb{E}X_i = a_i$, $\mathbb{E}X = a$. Write $S_i = X - X_i$, so that S_i and X_i are independent, and also take S_i and X_i^* to be independent; this is used to obtain the final inequality in (10) below.

We have for all bounded functions g,

$$\begin{aligned}
\mathbb{E}g(X^*) &= \mathbb{E}(Xg(X))/a \\
&= \sum_{i=1}^n (a_i/a)\mathbb{E}(X_i g(S_i + X_i))/a_i \\
&= \sum_{i=1}^n (a_i/a)\mathbb{E}g(S_i + X_i^*).
\end{aligned} \tag{10}$$

The result in (10) says precisely that X^* can be represented by the mixture of the distributions $S_i + X_i^*$ with mixture probabilities a_i/a. In words, in order to size bias the sum X with independent summands, we first pick an independent index I with probability proportional to its expectation, that is, with distribution $\mathbb{P}(I = i) = a_i/a$, and then size bias only the summand X_I. Or, with $X_1, \ldots, X_n, X_1^*, \ldots, X_n^*$ and I all independent

$$(X_1 + X_2 + \cdots + X_n)^* = X_1 + \cdots + X_{I-1} + X_I^* + X_{I+1} + \cdots + X_n. \tag{11}$$

For the special case where the summands X_i are not only independent but also *identically distributed*, or i.i.d., this recipe simplifies. In this case it does not matter which summand is biased, as all the distributions in the mixture are the same; hence for any $i = 1, \ldots, n$, $X^* \overset{\mathrm{d}}{=} X_1 + \cdots + X_{i-1} + X_i^* + X_{i+1} + \cdots + X_n$. In particular we may use $i = 1$ so that

$$(X_1 + X_2 + \cdots + X_n)^* = X_1^* + X_2 + X_3 + \cdots + X_n. \qquad (12)$$

5 Waiting for a bus: the renewal theory connection

Renewal theory provides a conceptual explanation of the identity (12) and at the same time gives an explanation of the waiting time paradox. Let the interarrival times of our buses in Section 1 be denoted X_i, so that buses arrive at times $X_1, X_1 + X_2, X_1 + X_2 + X_3, \ldots$, and assume only that the X_i are i.i.d., strictly positive random variables with finite mean; the paradox presented earlier was the special case with X_i exponentially distributed. Implicit in the story of my arrival time T as “arbitrary” is that my precise arrival time does not matter, and that there should be no relation between my arrival time and the schedule of buses. One way to model this assumption is to choose T uniformly from 0 to l, independent of $X_1, X_2, \ldots$, and then take the limit as $l \to \infty$; informally, just imagine some very large l. Such a T corresponds to throwing a dart at random from a great distance toward the real line, which has been subdivided into intervals of lengths X_i. Naturally the dart is twice as likely to land in a given interval of length two than one of length one, and generally x times as likely to land in a given interval of length x as one of length one. In other words, if the interarrival times X_i have a distribution $dF(x)$, the distribution of the length of the interval where the dart lands is proportional to $x\ dF(x)$. The constant of proportionality must be $1/a$, in order to make a legitimate distribution, so the distribution of the interval where the dart lands is the distribution of X^*.

The conceptual explanation of identity (12) is the following. Suppose that every n^{th} bus is bright blue, so that the waiting time between bright blue buses is the sum over a block of n successive arrival times. Again, the random time T finds itself in an interval whose length is distributed as the size biased distribution of the interarrival times; the length of the neighboring intervals are not affected. But by considering the variables as appearing in blocks of n, the random time T must also find itself in a block distributed as $(X_1+\cdots+X_n)^*$. Since only the interval containing one of the interarrival times has been size biased, this sum must be equal in distribution to $X_1+\cdots+X_{i-1}+X_i^*+X_{i+1}+\cdots+X_n$.

A more precise explanation of our waiting time paradox is based on the concept of stationarity — randomizing the schedule of buses so that I can arrive at an arbitrary time t, and specifying a particular t does not influence how long I must wait for the next bus. The simple process with arrivals at times $X_1, X_1+X_2, X_1+X_2+X_3, \ldots$ is in general not stationary; and the distribution of the time W_t that we wait from time t until the arrival of the next bus varies with t. We can, however, cook up a stationary process from this simple process by a modification suggested by size biasing. For motivation, recall the case where I arrive at T chosen uniformly from $(0, l)$. In the limit as $l \to \infty$ the interval containing T has length distributed as X_i^*, and my arrival within this interval is 'completely random.' That is, I wait UX_i^* for the next bus, and I missed the previous bus by $(1-U)X_i^*$, where U is uniform on (0,1) and independent of X_i^*. Thus it is plausible that one can form a stationary renewal process by the following recipe. Extend $X_1, X_2, \ldots$ to an independent, identically distributed sequence $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$. Let X_0^* be the size biased version of X_0 and let U be chosen uniformly in (0,1), with all variables independent. The origin is to occupy an interval of length X_0^*, and the location of the origin is to be uniformly distributed over this interval; hence buses arrive at time UX_0^* and $-(1-U)X_0^*$. Using $X_1, X_2, \ldots$ and $X_{-1}, X_{-2}, \ldots$ as interarrival times on the positive and negative side, we obtain a process by setting bus arrivals at the positive times $UX_0^*, UX_0^*+X_1, UX_0^*+X_1+X_2, \cdots$, and at the negative times $-(1-U)X_0^*, -((1-U)X_0^*+X_{-1}), -((1-U)X_0^*+X_{-1}+X_{-2}), \ldots$, and it can be proved that this process is stationary.

The interval which covers the origin has expected length $\mathbb{E}X_0^* = \mathbb{E}X_0^2/\mathbb{E}X_0$ (by (4) with $n=1$,) and the ratio of this to $\mathbb{E}X_0$ is $\mathbb{E}X_0^*/\mathbb{E}X_0 = \mathbb{E}X_0^2/(\mathbb{E}X_0)^2$. By Cauchy-Schwarz, this ratio is at least 1; and every value in $[1,\infty]$ is feasible. Note that my waiting time is $\mathbb{E}W_T = \mathbb{E}W_0 = \mathbb{E}(UX_0^*) = (1/2)\mathbb{E}X_0^*$, so the ratio of my waiting time to the average time between buses can be any value between 1/2 and infinity, depending on the distribution of the interarrival times.

The exponential case is very special, where strange and wonderful “coincidences” effectively hide all the structure involved in size biasing and stationarity. The distribution of X_0^*, obtained by size biasing the unit exponential, has density xe^{-x} for $x > 0$, using (3) with $a = 1$. This distribution is known as Gamma(1,2). In particular, $\mathbb{E}X_i^* = \int_0^\infty x(xe^{-x})\ dx = 2$, and splitting this in half for “symmetry” as in Feller’s answer (b) gives 1 as the expected time I must wait for the next bus. Furthermore, the independent uniform U splits that Gamma(1,2) variable X_0^* into UX_0^* and $(1-U)X_0^*$, and these turn out to be independent, and each having the original exponential distribution. Thus the general recipe for cooking up a stationary process, involving X_0^* and U in general, simplifies beyond recognition: the original simple schedule with arrivals at times $X_1, X_1+X_2, X_1+X_2+X_3, \ldots$ forms half of a stationary process, which is completed by its other half, arrivals at $-X_1', -(X_1'+X_2'), \ldots$, with $X_1, X_2, \ldots, X_1', X_2', \ldots$ all independent and exponentially distributed.

6 Size bias in statistics

But size biasing is not always undesired. In fact, it can be used to construct *unbiased* estimators of quantities that are at first glance difficult to estimate without bias. Suppose we have a population of n individuals, and associated to each individual i is the pair of real numbers $x_i \geq 0$ and y_i, with $\sum x_i > 0$. Perhaps x_i is how much the i^{th} customer was billed by their utility company last month, and y_i, say a smaller value than x_i, the amount they were supposed to have been billed. Suppose we would like to know just how severe the overbilling error is; we would like to know the ’adjustment factor’, which is the ratio $\sum_i y_i / \sum_i x_i$. Collecting the paired values for everyone is laborious and expensive, so we would like to be able to use a sample of $m < n$ pairs to make an estimate. It is not too hard to verify that if we choose a set R by selecting m pairs uniformly from the n, then the estimate $\sum_{j\in R} y_j / \sum_{j\in R} x_j$ will be biased; that is, the estimate, on average, will not equal the ratio we are trying to estimate.

Here’s how size biasing can be used to construct an unbiased estimate of the ratio $\sum_i y_i / \sum_i x_i$, using $m < n$ pairs. Create a random set $\tilde{R}$ of size m by first selecting a pair with probability proportional to x_i, and then $m-1$ pairs uniformly from the remaining pairs. Though we are out of the independent framework, the principle of (12) is still at work; size biasing one has size biased the sum. (This is so because we have size biased the one, and then chosen the others from an appropriate conditional distribution.) That is, one can now show that by biasing to include the single element in proportion to its x value, we have achieved a distribution whereby the probability of choosing the set r is proportional to $\sum_{j \in r} x_j$. From this observation it is not hard to see why $\mathbb{E}(\sum_{j \in \tilde{R}} y_j / \sum_{j \in \tilde{R}} x_j) = \sum_i y_i / \sum_i x_i$. This method is known as Midzuno's procedure for unbiased ratio estimation, and is noted in Cochran [8].

7 Size biasing, tightness, and uniform integrability

Recall that a collection of random variables $\{Y_\alpha : \alpha \in I\}$ is *tight* iff for all $\varepsilon > 0$ there exists $L < \infty$ such that

$$\mathbb{P}(Y_\alpha \notin [-L, L]) < \varepsilon \quad \text{for all } \alpha \in I.$$

This definition looks quite similar to the definition of uniform integrability, where we say $\{X_\alpha : \alpha \in I\}$ is *uniformly integrable*, or UI, iff for all $\delta > 0$ there exists $L < \infty$ such that

$$\mathbb{E}(|X_\alpha|; X_\alpha \notin [-L, L]) < \delta \quad \text{for all } \alpha \in I.$$

Intuitively, tightness for a family is that uniformly over the family, the probability mass due to large values is arbitrarily small. Similarly, uniform integrability is the condition that, uniformly over the family, the contribution to the expectation due to large values is arbitrarily small.

Tightness of the family of random variables $\{Y_\alpha : \alpha \in I\}$ implies that every sequence of variables $Y_n, n = 1, 2, \ldots$ from the family has a subsequence that converges in distribution. The concept of tightness is very useful not just for random variables, that is, real-valued random objects, but also for random elements of other spaces; in more general spaces, the closed intervals $[-L, L]$ are replaced by *compact sets*. If $\{X_\alpha : \alpha \in I\}$ is uniformly integrable, $\mathbb{E}X_n \to \mathbb{E}X$ for any sequence of variables $X_n, n = 1, 2, \ldots$ from the family that converges in distribution.

To discuss the connection between size biasing and uniform integrability, it is useful to restate the basic definitions in terms of nonnegative random variables. It is clear from the definition of tightness above that a family of *nonnegative* random variables $\{Y_\alpha : \alpha \in I\}$ is tight iff for all $\varepsilon > 0$ there exists $L < \infty$ such that

$$\mathbb{P}(Y_\alpha > L) < \varepsilon \quad \text{for all } \alpha \in I, \tag{13}$$

and from the definition of UI, that a family of *nonnegative* random variables $\{X_\alpha : \alpha \in I\}$ is uniformly integrable iff for all $\delta > 0$ there exists $L < \infty$ such that

$$\mathbb{E}(X_\alpha; X_\alpha > L) < \delta \quad \text{for all } \alpha \in I. \tag{14}$$

For general random variables, the family $\{G_\alpha : \alpha \in I\}$ is tight [respectively UI] iff $\{|G_\alpha| : \alpha \in I\}$ is tight [respectively UI]. We specialize in the remainder of this section to random variables that are non-negative with finite, strictly positive mean.

Since *size bias* relates contribution to the expectation to probability mass, there should be a connection between tightness, size bias, and UI. However, care should be taken to distinguish between the (additive) contribution to expectation, and the *relative* contribution to expectation. The following example makes this distinction clear. Let

$$\mathbb{P}(X_n = n) = 1/n^2, \mathbb{P}(X_n = 0) = 1 - 1/n^2, \quad n = 1, 2, \ldots.$$

Here, $\mathbb{E}X_n = 1/n$, the family $\{X_n\}$ is uniformly integrable, but $1 = \mathbb{P}(X_n^* = n)$, so the family $\{X_n^*\}$ is not tight. The trouble is that the additive contribution to the expectation from large values of X_n is small, but the *relative* contribution is large — one hundred percent! The following two theorems, which exclude this phenomenon, show that tightness and uniform integrability are very closely related.

Theorem 7.1 *Assume that for $\alpha \in I$, where I is an arbitrary index set, the random variables X_α satisfy $X_\alpha \geq 0$ and $c \leq \mathbb{E}X_\alpha < \infty$, for some $c > 0$. For each α let $Y_\alpha = X_\alpha^*$. Then*

$$\{X_\alpha : \alpha \in I\} \textit{ is UI iff } \{Y_\alpha : \alpha \in I\} \textit{ is tight.}$$

Proof. First, with $Y_\alpha = X_\alpha^*$, we have $\mathbb{P}(Y_\alpha > L) = \mathbb{E}(1(Y_\alpha > L)) = \mathbb{E}(X_\alpha 1(X_\alpha > L))/\mathbb{E}X_\alpha$, so for any L and $\alpha \in I$,

$$\mathbb{E}(X_\alpha; X_\alpha > L) = \mathbb{E}X_\alpha \mathbb{P}(Y_\alpha > L).$$

Assume that $\{X_\alpha : \alpha \in I\}$ is UI, and let $\varepsilon > 0$ be given to test tightness in (13). Let L be such that (14) is satisfied with $\delta = \varepsilon c$. Now, using $\mathbb{E}X_\alpha \geq c$, for every $\alpha \in I$,

$$\mathbb{P}(Y_\alpha > L) = \mathbb{E}(X_\alpha; X_\alpha > L)/\mathbb{E}X_\alpha \leq \mathbb{E}(X_\alpha; X_\alpha > L)/c < \delta/c = \varepsilon,$$

establishing (13).

Second, assume that $\{X_\alpha : \alpha \in I\}$ if tight, and take L_0 to satisfy (13) with $\varepsilon := 1/2$, so that $\mathbb{P}(Y_\alpha > L_0) < 1/2$ for all $\alpha \in I$. Hence, for all $\alpha \in I$,

$$\mathbb{E}(X_\alpha; X_\alpha > L_0) = \mathbb{E}X_\alpha \mathbb{P}(Y_\alpha > L_0) < \mathbb{E}X_\alpha/2,$$

and therefore,

$$\begin{aligned} L_0 \geq \mathbb{E}(X_\alpha; X_\alpha \leq L_0) &= \mathbb{E}X_\alpha - \mathbb{E}(X_\alpha; X_\alpha > L_0) \\ &> \mathbb{E}X_\alpha - \mathbb{E}X_\alpha/2 = \mathbb{E}X_\alpha/2, \end{aligned}$$

and hence $\mathbb{E}X_\alpha < 2L_0$. Now given $\delta > 0$ let L satisfy (13) for $\varepsilon = \delta/(2L_0)$. Hence $\forall \alpha \in I$,

$$\mathbb{E}(X_\alpha; X_\alpha > L) = \mathbb{E}X_\alpha \ \mathbb{P}(Y_\alpha > L) < 2L_0 \ \mathbb{P}(Y_\alpha > L) < 2L_0 \ \varepsilon = \delta,$$

establishing (14).

Theorem 7.2 *Assume the for $\alpha \in I$, where I is an arbitrary index set, that random variables X_α satisfy $X_\alpha \geq 0$ and $\mathbb{E}X_\alpha < \infty$. Pick any $c \in (0, \infty)$, and for each α let $Y_\alpha = (c + X_\alpha)^*$. Then*

$$\{X_\alpha : \alpha \in I\} \text{ is UI iff } \{Y_\alpha : \alpha \in I\} \text{ is tight.}$$

Proof. By Theorem 7.1, the family $\{c + X_\alpha\}$ is UI iff the family $\{(c + X_\alpha)^*\}$ is tight. As it is easy to verify that the family $\{X_\alpha\}$ is tight [respectively UI] iff the family $\{c+X_\alpha\}$ is tight [respectively UI], Theorem 7.2 follows directly from Theorem 7.1.

8 Size biasing and infinite divisibility: the heuristic

Because of the recipe (12), it is natural that our question in (5) is related to the concept of infinite divisibility. We say that a random variable X is infinitely divisible if for all n, X can be decomposed in distribution as the sum of n iid variables. That is, that for all n there exists a distribution dF_n such that if $X_1^{(n)}, \ldots, X_n^{(n)}$ are iid with this distribution, then

$$X \overset{\mathrm{d}}{=} X_1^{(n)} + \cdots + X_n^{(n)}. \tag{15}$$

Because this is an iid sum, by (12), we have

$$X^* = (X - X_1^{(n)}) \ + \ (X_1^{(n)})^*,$$

with $X - X_1^{(n)}$ and $(X_1^{(n)})^*$ independent. For large n, $X - X_1^{(n)}$ will be close to X, and so we have represented the size bias distribution of X as approximately equal, in distribution, to X plus an independent increment. Hence it is natural to suspect that the class of non negative infinitely divisible random variables can be size biased by adding an independent increment.

It is not difficult to make the above argument rigorous, for infinitely divisible $X \geq 0$ with $\mathbb{E}X < \infty$. First, to show that $X - X_1^{(n)}$ converges in distribution to X, it suffices to show that $X_1^{(n)}$ converges to zero in probability. Note that $X \geq 0$ implies $X_1^{(n)} \geq 0$, since (15) gives $0 = \mathbb{P}(X < 0) \geq (\mathbb{P}(X_1^{(n)} < 0)^n$. Then, given $\epsilon > 0$, $\infty > \mathbb{E}X \geq n\mathbb{P}(X_1^{(n)} > \epsilon)\epsilon$ implies that $\mathbb{P}(X_1^{(n)} > \epsilon) \to 0$; hence $X_1^{(n)} \to 0$ in probability as $n \to \infty$.

We have that

$$X^* = (X - X_1^{(n)}) + (X_1^{(n)})^*, \tag{16}$$

with $X - X_1^{(n)}$ and $(X_1^{(n)})^*$ independent, and $X - X_1^{(n)}$ converging to X in distribution. Now, the family of random variables $(X_1^{(n)})^*$ is "tight", because given $\epsilon > 0$, there is a K such that $\mathbb{P}(X^* > K) < \epsilon$, and by (16), for all n, $\mathbb{P}((X_1^{(n)})^* > K) \leq \mathbb{P}(X^* > K) < \epsilon$. Thus, by Helly's theorem, there exists a subsequence n_k of the n's along which $(X_1^{(n)})^*$ converges in distribution, say $(X_1^{(n_k)})^* \overset{\text{distr}}{\longrightarrow} Y$. Taking $n \to \infty$ along this subsequence, the pair $(X - X_1^n, (X_1^n)^*)$ converges jointly to the pair (X, Y) with X and Y independent. From $X^* = (X - X_1^{(n_k)}) + (X_1^{(n_k)})^* \overset{\text{distr}}{\longrightarrow} X + Y$ as $k \to \infty$ we conclude that $X^* \overset{\text{d}}{=} X + Y$, with $Y \geq 0$, and X, Y independent. This concludes a proof that if $X \geq 0$ with $0 < \mathbb{E}X < \infty$ is infinitely divisible, then it satisfies (5).

9 Size biasing and Compound Poisson

Let us now return to our main theme, determining for which distributions we have (5). We have already seen it is true, trivially, for a scale multiple of a Poisson random variable. We combine this with the observation (11) that to size bias a sum of independent random variables, just bias a single summand, chosen proportional to its expectation. Consider a random variable of the form

$$X = \sum_{1}^{n} X_j, \quad \text{with } X_j = y_j Z_j, \quad Z_j \sim Po(\lambda_j), \quad Z_1, \ldots, Z_n \, \text{independent}, \tag{17}$$

with distinct constants $y_j > 0$.

Since X is a sum of independent variables, we can size bias X by the recipe (11); pick a summand proportional to its expectation and size bias that one. We have $EX_j = y_j \lambda_j$ and therefore $a = EX = \sum_j y_j \lambda_j$. Hence, the probability that we pick summand j to size bias is

$$\mathbb{P}(I = j) = y_j \lambda_j / a.$$

But by (9), $X_j^* = X_j + y_j$, so that when we pick X_j to bias we add y_j. Hence, to bias X we merely add y_j with probability $y_j \lambda_j / a$, or, to put it another way $X^* = X + Y$, with X, Y independent and

$$\mathbb{P}(Y = y_j) = y_j \lambda_j / a. \tag{18}$$

In summary, X of the form (17) can be size biased by adding an independent, nonnegative increment. It will turn out that we have now nearly found all solutions of (5), which will be obtained by taking limits of variables type (17) and adding a nonnegative constant.

Sums of the form (17) are said to have a compound Poisson distribution — of finite type. Compound Poisson variables in general are obtained by a method which at first glance looks unrelated to (17), considering the sum S_N formed by adding a Poisson number N of iid summands $A_1, A_2, \ldots$ from *any* distribution, i.e. taking

$$X \overset{\mathrm{d}}{=} S_N := A_1 + \cdots + A_N, \quad N \sim Po(\lambda), \tag{19}$$

where $N, A_1, A_2, \ldots$ are independent and $A_1, A_2, \ldots$ identically distributed. The notation $S_N := A_1 + \ldots + A_N$ reflects that of a random walk, $S_n := A_1 + \cdots A_n$ for $n = 0, 1, 2, \ldots$, with $S_0 = 0$.

To fit the sum (17) into the form (19), let $\lambda = \sum \lambda_j$ and let $A, A_1, A_2, \ldots$ be iid with

$$\mathbb{P}(A = y_j) = \lambda_j / \lambda. \tag{20}$$

The claim is that X as specified by (17) has the same distribution as the sum S_N. Checking this claim is most easily done with generating functions, such as characteristic functions, discussed in the next section. Nevertheless, it is an enjoyable exercise for the special case where $a_1, \ldots, a_n$ are mutually irrational, i.e. linearly independent over the rationals, so that with $k_1, \ldots, k_n$ integers, the sum $\sum_1^n k_j a_j$ determines the values of the k_j.

Note that the distribution (20) of the summands A_i is *different* from the distribution in (18). In fact, $A^* \overset{\mathrm{d}}{=} Y$, which can be checked using (20) together with (3), and comparing the result to (18):

$$\mathbb{P}(A^* = y_j) = \frac{y_j \mathbb{P}(A = y_j)}{\mathbb{E}A} = \frac{y_j \lambda_j / \lambda}{\sum_k y_k \lambda_k / \lambda} = \frac{y_j \lambda_j}{\sum_k y_k \lambda_k} = \mathbb{P}(Y = y_j). \tag{21}$$

Thus the result that for the compound Poisson of finite type, $X^* = X + Y$ can be expressed as

$$(S_N)^* = S_N + Y \overset{\mathrm{d}}{=} A_1 + \cdots + A_N + A^*, \tag{22}$$

with all summands independent. Note how this contrasts with the recipe (12) for size biasing a sum with a deterministic number of terms n; in the case of (12) the biased sum and the original sum have the same number of terms, but in (22) the biased sum has one more term than the original sum.

If we want to size bias a general compound Poisson random variable S_N, there must be some restrictions on the distribution for the iid summands A_i in (19). First, since $\lambda > 0$, and (using the independence of N and $S_0, S_1, \ldots$,) $\mathbb{E}S_N = \mathbb{E}N \; \mathbb{E}A = \lambda \mathbb{E}A$, the condition that $\mathbb{E}S_N \in (0, \infty)$ is equivalent to $\mathbb{E}A \in (0, \infty)$. The condition that $S_N \geq 0$ is equivalent to the condition $A_i \geq 0$. We *choose* the additional requirement that A be *strictly* positive, which is convenient since it enables the simple computation $\mathbb{P}(S_N = 0) = \mathbb{P}(N = 0) = e^{-\lambda}$. There is no loss of generality in this added restriction, for if $p := \mathbb{P}(A = 0) > 0$, then with M being Poisson with parameter $\mathbb{E}M = \lambda(1 - p)$, and $B_1, B_2, \ldots$ iid and independent of M, with $\mathbb{P}(B \in I) = \mathbb{P}(A \in I)/(1 - p)$ for $I \subset (0, \infty)$ (using $p < 1$ since $\mathbb{E}A > 0$), we have $A_1 + \cdots + A_N \overset{\mathrm{d}}{=} B_1 + \cdots + B_M$, so that S_N can be represented as a compound Poisson with *strictly* positive summands.

10 Compound Poisson vs. Infinitely Divisible

We now have two ways to produce solutions of (5), infinitely divisible distributions from section 8, and finite type compound Poisson distributions, from (17), so naturally the next question is: how are these related?

The finite type compound Poisson random variable in (17) can be extended by adding in a nonnegative constant c, to get X of the form

$$X = c + \sum_1^n X_j, \quad \text{with } X_j = y_j Z_j, \quad Z_j \sim Po(\lambda_j), \quad Z_1, \ldots, Z_n \text{ independent,} \tag{23}$$

with $c \geq 0$ and distinct constants $y_j > 0$. In case $c > 0$, this random variable is not called compound Poisson. Every random variable of the form (23) is infinitely divisible – for the $X_i^{(m)}$ [for $i = 1$ to m in (15)] simply take a sum of the same form, but with c replaced by c/m and λ_j replaced by λ_j/m.

The following two facts are *not* meant to be obvious. By taking distributional limits of sums of the form (23), and requiring that $c + \sum_1^n y_j \lambda_j$ stays bounded as $n \to \infty$, one gets all the non-negative infinitely divisible distributions with finite mean. By also requiring that $c = 0$ and $\sum_1^n \lambda_j$ stays bounded, the limits are all the finite mean, non-negative compound Poisson distributions.

To proceed, we calculate the characteristic function for the distribution in (17). First, if X is Po(λ), then

$$\phi_X(u) := \mathbb{E} e^{iuX} = \sum_{k \geq 0} e^{iuk} \mathbb{P}(X = k) = \sum e^{iuk} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum \frac{(\lambda e^{iu})^k}{k!} = \exp(\lambda(e^{iu} - 1)).$$

For a scalar multiple yX of a Poisson random variable,

$$\phi_{yX}(u) = \mathbb{E} e^{iu(yX)} = \mathbb{E} e^{i(yu)X} = \phi_X(yu) = \exp(\lambda(e^{iuy} - 1)).$$

Thus the summand $X_j = y_j Z_j$ in (17) has characteristic function $\exp(\lambda_j(e^{iuy_j} - 1))$, and hence the sum X has characteristic function

$$\phi_X(u) = \prod_{j=1}^n \exp\left(\lambda_j(e^{iuy_j} - 1)\right) = \exp\left(\sum_{j=1}^n \lambda_j (e^{iuy_j} - 1)\right).$$

To prepare for taking limits, we write this as

$$\phi_X(u) = \exp\left(\sum_1^n \lambda_j(e^{iuy_j} - 1)\right) = \exp\left(\int_{(0,\infty)} (e^{iuy} - 1)\ \mu(dy)\right), \tag{24}$$

where μ is the measure on $(0,\infty)$ which places mass λ_j at location y_j. The total mass of μ is $\int 1\ \mu(dy) = \sum_1^n \lambda_j$, which we will denote by λ, and the first moment of μ is $\int y\ \mu(dy) = \sum_1^n y_j\lambda_j$, which happens to equal $a := \mathbb{E}X$.

Allowing the addition of a constant c, the random variable of the form (23) has characteristic function ϕ_X whose logarithm has the form $\log \phi_X(u) = iuc + \int_{(0,\infty)}(e^{iuy} - 1)\ \mu(dy)$, where μ is a measure whose support consists of a finite number of points. The finite mean, not identically zero distributional limits of such random variables yield all of the finite mean, nonnegative, not identically zero, infinitely divisible distributions. A random variable X with such a distribution has characteristic function ϕ_X with

$$\log \phi_X(u) = iuc + \int_{(0,\infty)} (e^{iuy} - 1)\ \mu(dy), \tag{25}$$

where $c \geq 0$, μ is any nonnegative measure on $(0,\infty)$ such that $\int y\ \mu(dy) < \infty$, and not both c and μ are zero.

Which of the distributions above are compound Poisson? The compound Poisson variables are the ones in which $c = 0$ and $\lambda := \mu((0,\infty)) < \infty$. With this choice of λ for the parameter of N , we have $X \overset{\mathrm{d}}{=} S_N$ as in (19), with the distribution of A given by μ/λ, i.e. $\mathbb{P}(A \in dy) = \mu(dy)/\lambda$. To check, note that for $X = S_N$, $\phi_X(u) := \mathbb{E}e^{iuS_N} = \sum_{n\geq 0} \mathbb{P}(N = n)e^{iuS_n} = e^{-\lambda}\sum_{n\geq 0} \lambda^n/n!(\phi_A(u))^n = \exp(\lambda(\phi_A(u) - 1)$ $= \exp(\lambda(\int_{(0,\infty)}(e^{iuy} - 1)\ \mathbb{P}(A \in dy)$, which is exactly (25) with $c = 0$.

In our context, it is easy to tell whether or not a given infinitely divisible random variable is also compound Poisson — it is if and only if $\mathbb{P}(X = 0)$ is strictly positive — corresponding to $\mathbb{P}(N = 0) = e^{-\lambda}$ and $e^{-\infty} = 0$. Among the examples of infinitely divisible random variables in section 16.2, the only compound Poisson examples are the Geometric and Negative binomial family, and the distributions related to Buchstab's function.

In (18) — specifying the distribution of the increment Y when $X^* = X + Y$ with X, Y independent — the factor y_j on the right hand side suggests another way to write (25). We multiply and divide by y to get

$$\log \phi_X(u) = \int_{[0,\infty)} \frac{e^{iuy}-1}{y}\ \nu(dy), \tag{26}$$

where ν is any nonnegative measure on $[0,\infty)$ with total mass $\nu(\ [0,\infty)\ ) \in (0,\infty)$. The measure ν on $[0,\infty)$ is related to c and μ by $\nu(\{0\}) = c$ and for $y > 0,\ \ \nu(dy) = y\ \mu(dy)$; we follow the natural convention that $(e^{iuy}-1)/y$ for $y = 0$ is interpreted as iu. The measure ν/a is a probability measure on $[0,\infty)$ because

$$a := \mathbb{E}X = -i(d\log\phi_X(u)/du)|_{u=0} = c + \int_{(0,\infty)} y\ \mu(dy) = c + \nu((0,\infty)) = \nu([0,\infty)). \tag{27}$$

We believe it is proper to refer to either (25) or (26) as a Lévy representation, and to refer to either μ or ν as the Lévy measure, in honor of Paul Lévy; indeed when restriction that X be nonnegative is dropped, there are still more forms for the Lévy representation, see e.g. [9] Feller volume II, chapter XVII.

It is not hard to see that any distribution specified by (26) satisfies (5), by the following calculation with characteristic functions. Note that for $g(x) := e^{iux}$, the characterization (2) of the distribution of X^* directly gives the characteristic function ϕ^* of X^* as $\phi^*(u) := \mathbb{E}e^{iuX^*} = \mathbb{E}(Xe^{iuX})/a$. For any $X \geq 0$ with finite mean, by an application of the dominated convergence theorem, if $\phi(u) := \mathbb{E}e^{iuX}$ then $\phi'(u) = \mathbb{E}(iXe^{iuX})$. Thus for any $X \geq 0$ with $0 < a = \mathbb{E}X < \infty$,

$$\phi^*(u) = \frac{1}{ia}\ \phi'(u). \tag{28}$$

Now if X has characteristic function ϕ given by (26), again using dominated convergence,

$$\phi'(u) = \phi(u)\ \left(ic + \int_{(0,\infty)} iye^{iuy}\ \mu(dy)\right) = ia\ \phi(u)\ \int_{[0,\infty)} e^{iuy}\nu(dy)/a. \tag{29}$$

Taking the probability measure ν/a as the distribution of Y, and writing η for the characteristic function of Y, (29) says that $\phi'(u) = ia\ \phi(u)\ \eta(u)$. Combined with (28), we have

$$\phi^* = \phi\ \eta. \tag{30}$$

Thus $X^* = X + Y$, with X and Y independent and $\mathcal{L}(Y) = \nu/a$.

For the compound Poisson case in general, in which the distribution of A is μ/λ, we have $Y \overset{\mathrm{d}}{=} A^*$ because $a := \mathbb{E}X = \lambda \, \mathbb{E}A$ and

$$\mathbb{P}(A^* \in dy) = y\mathbb{P}(A \in dy)/\mathbb{E}A = \frac{y(\mu(dy)/\lambda)}{a/\lambda} = \nu(dy)/a = \mathbb{P}(Y \in dy), \tag{31}$$

which can be compared discrete version (20). Thus the computation (30) shows, for the compound Poisson case, that (22) holds.

11 Main Result

Our main result is essentially the converse of the computation (30).

Theorem 11.1 *(Steutel 1973 [17]) For a random variable $X \geq 0$ with $a := \mathbb{E}X \in (0, \infty)$, the following three statements are equivalent.*

i) There exists a coupling with $X^* = X + Y, Y \geq 0$, and X, Y independent.

ii) The distribution of X is infinitely divisible,

iii) The characteristic function of X has the Lévy representation (26).

Furthermore, when any of these statements hold, the Lévy measure ν in (26) equals a times the distribution of Y.

Proof. We have proved that ii) implies i), in section 8. We have proved that iii) implies i), in the argument ending with (30), which also shows that given ν in the Lévy representation (26), the increment Y in i) has distribution ν/a.

The equivalence of ii) and iii) is a standard textbook topic — with the argument for iii) implies ii) being simply that X with a given ν is the sum of n iid terms each having the Lévy representation (26) with ν/n playing the role of ν.

Now to prove that i) implies ii), we assume that i) holds. The characteristic function ϕ^* of X^* has the form $\phi^* = \phi \, \eta$, where ϕ and η are the characteristic functions of X and Y, so that $\eta(u) = \mathbb{E}e^{iuY} = \int_{[0,\infty)} e^{iuy} \, \mathbb{P}(Y \in dy)$. Combining this with (28) we have

$$\frac{1}{ia} \, \phi'(u) = \phi^*(u) = \phi(u) \; \eta(u)$$

so that $(\log \phi(u))' = ia\ \eta(u)$. Since $\log \phi(0) = 0$,

$$\begin{aligned}
\log \phi(u) &= ia \int_{s\in[0,u)} \eta(s)\ ds \\
&= ia \int_{s\in[0,u)} \int_{y\in[0,\infty)} e^{isy}\ \mathbb{P}(Y \in dy)\ ds \\
&= ia \int_{y\in[0,\infty)} \int_{s\in[0,u)} e^{isy}\ ds\ \mathbb{P}(Y \in dy) \\
&= ia \left(\int_{y\in(0,\infty)} \int_{s\in[0,u)} e^{isy}\ ds\ \mathbb{P}(Y \in dy)\ + u\mathbb{P}(Y = 0) \right) \\
&= a \int_{y\in[0,\infty)} \frac{e^{iuy} - 1}{y}\ \mathbb{P}(Y \in dy).
\end{aligned}$$

This is the same as the representation (26), with $\nu = a\mathcal{L}(Y)$ for the random variable Y given in i).

■

Observe that ν is an arbitrary probability distribution on $[0, \infty)$, i.e. $\nu \in$Pr$([0, \infty))$, and the choice of $a \in (0, \infty)$ is also arbitrary. Thus there is a one-to-one correspondence between the Cartesian product Pr$([0, \infty)) \times (0, \infty)$ and the set of the nonnegative, infinitely divisible distributions with finite, strictly positive mean.

12 A consequence of $X^* = X + Y$ with independence

To paraphrase the result of Theorem 11.1, for a nonnegative random variable X with $a := \mathbb{E}X \in (0, \infty)$, it is possible to find a coupling with $X^* = X + Y$, $Y \geq 0$ and X, Y independent *if and only if* the distribution of X is the infinitely divisible distribution with Lévy representation (26) governed by the finite measure ν equal to a times the distribution of Y. Thus we know an explicit, albeit complicated, relation between the distributions of X and Y. It is worth seeing how (5) directly gives a simple relation between the densities of X and Y, if these densities exist.

In the discrete case, if X has a mass function f_X and if (26) holds, then Y must mass function, f_Y, and by (5), f_{X^*} is the convolution of f_X and f_Y: $f_{X^*}(x) = \sum_y f_X(x - y) f_Y(y)$. Combined with (3), this says that for all $x > 0$,

$$f_X(x) = \frac{a}{x} \sum_y f_X(x-y) f_Y(y).$$

Likewise, in the continuous case, if X has density f_X (i.e. if for all bounded g, $\mathbb{E}g(X) = \int g(x) f_X(x)\, dx$,) and if (26) holds, *and if further* Y has a density f_Y, then by (5), f_{X^*} is the convolution of f_X and f_Y: $f_{X^*}(x) = \int_y f_X(x-y) f_Y(y)$. Combined with (3), this says that for all $x > 0$,

$$f_X(x) = \frac{a}{x} \int_y f_X(x-y) f_Y(y)\ dy. \tag{32}$$

13 Historical remark

For all intents and purposes, Theorem 11.1 is due to Steutel [17]. The way he states his result is sufficiently different from our Theorem 11.1 that for comparison, we quote verbatim from [17], p. 136:
Theorem 5.3. A d.f. F on $[0,\infty)$ is infinitely divisible iff it satisfies

$$(5.6) \qquad \int_0^x u\ dF(u) \ = \ \int_0^x F(x-u)\ dK(u),$$

where K is non-decreasing.

Observe that Steutel's result is actually more general than Theorem 11.1, since that latter only deals with nonnegative infinitely divisible random variables with *finite mean.* The explicit connection between the independent increment for size biasing, and the Lévy representation, is made in [11], along with further connections between renewal theory and independent increments.

14 The product rule for size biasing

We have seen that for independent, nonnegative random variables $X_1, \ldots, X_n$, the sum $X = X_1 + X_2 \cdots + X_n$ can be size biased by picking a single summand at random with probability proportional to its expectation, and replacing it with one from its size biased distribution. Is there a comparable procedure for the product $W = X_1 X_2 \cdots X_n$? Would it involve size-biasing a single factor?

Let $a_i = \mathbb{E}X_i \in (0,\infty)$, let F_i be the distribution function of X_i, and let F_i^* be the distribution function of X_i^*, so that $dF_i^*(x) = x\,dF_i(x)/a_i$. Let $X_1^*, \ldots, X_n^*$ be independent. By (2) with $a := \mathbb{E}W = a_1 a_2 \cdots a_n$, for all bounded functions g,

$$\begin{aligned}
Eg(W^*) &= E\,(Wg(W))\,/(a_1a_2\cdots a_n) \\
&= \int\cdots\int x_1\cdots x_n \;\; g(x_1x_2\cdots x_n)\; dF_1(x_1)\cdots dF_n(x_n)/(a_1\cdots a_n) \\
&= \int\cdots\int g(x_1x_2\cdots x_n)\;(x_1\; dF_1(x_1)/a_1)\cdots(x_n\; dF_n(x_n)/a_n) \\
&= \int\cdots\int g(x_1x_2\cdots x_n)\; dF_1^*(x_1)\cdots\;\; dF_n^*(x_n) \\
&= Eg(X_1^*\cdots X_n^*),
\end{aligned}$$

and so

$$W^* \overset{\mathrm{d}}{=} X_1^* \cdots X_n^*.$$

We have shown that to size bias a product of independent variables, one must size bias every factor making up the product, very much unlike what happens for a sum, where only one term is size biased!

15 Size biasing the lognormal distribution

The lognormal distribution is often used in financial mathematics to model prices, salaries, or values. A variable L with the lognormal distribution is obtained by exponentiating a normal variable. We follow the convention that Z denotes a standard normal, with $\mathbb{E}Z = 0$, var $Z = 1$, so that $L = e^Z$ represents a standard lognormal. With constants $\sigma > 0, \mu \in \mathbb{R}$, $\sigma Z + \mu$ represents the general normal, and $L = e^{\sigma Z + \mu}$ represents the general lognormal. As the lognormal is non-negative and has finite mean, it can be size biased to form L^*.

One way to guess the identity of L^* is to use the method of moments. For the standard case $L = e^Z$, for any real t, calculation gives $\mathbb{E}e^{tZ} = \exp(t^2/2)$. Taking $t = 1$ shows that $\mathbb{E}L = \sqrt{e}$, and more generally, for $n = 1, 2, \ldots$, $\mathbb{E}L^n = \mathbb{E}e^{nZ} = \exp(n^2/2)$. Using relation (4), the moment-shift for size biasing, we have $\mathbb{E}(L^*)^n = \mathbb{E}L^{n+1}/\mathbb{E}L = \exp((n+1)^2/2 - 1/2) = \exp(n^2/2 + n) = e^n\mathbb{E}L^n = \mathbb{E}(eL)^n$. Clearly we should guess that $L^* \overset{\mathrm{d}}{=} eL$, but we must be cautions, as the most famous example of a distribution which has moments of all orders but which is not determined by them is the lognormal; for other such distributions related to the normal, see [16].

We present a rigorous method for finding the distribution of L^*, based on the size biasing product rule of the previous section; as an exercise the reader might try to verify our conclusion (33) by working out the densities for lognormal distributions, and using the relation (3).

We begin with the case $\mu = 0, \sigma > 0$. Let C_i be independent variables taking the values 1 or -1 with equal probability. These variables have mean zero and variance one, and by the central limit theorem, we know that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma C_i \overset{\text{distr}}{\longrightarrow} \sigma Z.$$

Hence, we must have

$$W = \prod_{i=1}^{n} \exp(\frac{1}{\sqrt{n}} \sigma C_i) = \exp(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma C_i) \overset{\text{distr}}{\longrightarrow} \exp(\sigma Z) = L,$$

a lognormal, and thus $W^* \overset{\text{distr}}{\longrightarrow} L^*$. Write $X_i := \exp(\sigma C_i/\sqrt{n})$, so that $W = X_1 \cdots X_n$ with independent factors, and by the product rule, $W^* = X_1^* \cdots X_n^*$. The variables X_i take on the values $q = e^{-\sigma/\sqrt{n}}$ and $p = e^{\sigma/\sqrt{n}}$ with equal probability, and so X_i^* take on these same values, but with probabilities $q/(p+q)$ and $p/(p+q)$ respectively. Let's say that B_n of the X_i^* take the value p, so that $n - B_n$ of the X_i^* take the value q. Using B_n, we can write

$$W^* = p^{B_n} q^{n-B_n} = e^{\sigma(2B_n - n)/\sqrt{n}}.$$

Since B_n counts the number of "successes" in n independent trials, with success probability $p/(p+q)$, B_n is distributed binomial$(n, p/(p+q))$. As $n \to \infty$, the central limit theorem gives that B_n has an approximate normal distribution. Doing a second order Taylor expansion of e^x around zero, and applying it at $x = \pm\sigma/\sqrt{n}$, we find that $p/(p+q) = 1/2 + \sigma/(2\sqrt{n}) + O(1/n)$, so that B_n is approximately normal, with mean $np/(p+q) = (1/2)(n + \sigma\sqrt{n}) + O(1)$ and variance $npq/(p+q)^2 = n/4 + O(1/n^{3/2})$. Hence

$$\frac{1}{\sqrt{n}}(2B_n - n) \overset{\text{distr}}{\longrightarrow} Z + \sigma \quad \text{as } n \to \infty$$

and therefore

$$W^* \overset{\text{distr}}{\longrightarrow} e^{\sigma(Z+\sigma)}.$$

Since $W^* \overset{\text{distr}}{\longrightarrow} L^* = (e^{\sigma Z})^*$, we have shown that $(e^{\sigma Z})^* \overset{\text{d}}{=} e^{\sigma(Z+\sigma)}$. For the case where $L = e^{\sigma Z+\mu}$, the scaling relation (8) yields the formula for size biasing the lognormal in general:

$$(e^{\sigma Z+\mu})^* = e^{\sigma(Z+\sigma)+\mu}. \tag{33}$$

16 Examples

In light of Theorem 11.1, for a nonnegative random variable X with finite, strictly positive mean, being able to satisfy $X^* = X + Y$ with independence and $Y \geq 0$ is equivalent to being infinitely divisible. We give examples of size biasing, first with examples that are not infinitely divisible, then with examples that are.

16.1 Examples of size biasing without an independent increment

Both examples 1 and 2 below involve bounded, nonnegative random variables. Observe that in general, the distributions of X and X^* have the same support, except that always $\mathbb{P}(X^* = 0) = 0$. This immediately implies that if X is bounded but not constant, then it cannot satisfy (5).

Example 1. Bernoulli and binomial

Let B_i be Bernoulli with parameter $p \in (0,1]$, i.e. B_i takes the value 1 with probability p, and the value 0 with probability $1-p$. Clearly $B_i^* = 1$, since $\mathbb{P}(B_1^* = 1) = 1\mathbb{P}(B_1 = 1)/\mathbb{E}B_1 = 1$. If $B_1, B_2, \ldots$ are independent, and $S_n = B_1 + \cdots + B_n$ we say that $S_n \sim$ binomial (n,p). We size bias S_n by size biasing a single summand, so $S_n^* \overset{\text{d}}{=} S_{n-1} + 1$, which cannot be expressed as $S_n + Y$ with S_n, Y independent!

Note that letting $n \to \infty$ and $np \to \lambda$ in the relation $S_n^* \overset{\text{d}}{=} S_{n-1} + 1$ gives another proof that $X^* \overset{\text{d}}{=} X + 1$ when $X \sim Po(\lambda)$, because both $S_{n-1} \overset{\text{distr}}{\longrightarrow} X$ and $S_n \overset{\text{distr}}{\longrightarrow} X$. Here we have a family of examples without independence, whose limit is the basic example with independence.

Example 2. Uniform and Beta

The Beta distribution on (0,1), with parameters $a, b > 0$ is specified by saying that its has a density on (0,1), proportional to $(1-x)^{a-1}x^{b-1}$. The uniform distribution on (0,1) is the special case $a = b = 1$ of this Beta family. Using (3), if $X \sim \text{Beta}(a, b)$, then $X^* \sim \text{Beta}(a, b+1)$.

There are many families of distributions for which size biasing simply changes the parameters; our examples are the Beta family in example 2, the negative binomial family in example 4, the Gamma family in example 5, and the lognormal family in example 6. In these families, either all members satisfy (5), or else none do. Thus it might be tempting to guess that infinite divisibility is a property preserved by size biasing, but it ain't so.

Example 3. $X = 1 + W$ where W is Poisson

We have X^* is a mixture of $X + 0$ and $X + 1$, using (11) with $X_1 = 1$, $X_1^* = X_1 + 0$ and $X_2 = W$, $X_2^* = W + 1$. That is, X^* is a mixture of $1 + W$, with weight $1/(1 + \lambda)$, and $2 + W$, with weight $\lambda/(1+\lambda)$. Elementary calculation shows that it is not possible to have $X^* = X + Y$ with X, Y independent and $Y \geq 0$. Therefore X is not infinitely divisible.

Since $X = W^*$, we have an example in which W is infinitely divisible, but W^* is not.

16.2 Examples of $X^* = X + Y$ with independence

By Theorem 11.1, when X satisfies $X^* \overset{\text{d}}{=} X + Y$ with X, Y independent and $Y \geq 0$, the distribution of X is determined by the distribution of Y together with a choice for the constant $a \in (0, \infty)$ to serve as $\mathbb{E}X$. Thus all our examples below, organized by a choice of Y, come in one parameter families indexed by a — or if more convenient, by something proportional to a; in these families, X varies and Y stays constant!

Example 4. Y is 1+geometric. X is geometric or negative binomial

4a) The natural starting point is that you are given the geometric distribution: $\mathbb{P}(X = j) = (1 - q)q^j$ for $j \geq 0$, with $0 < q < 1$, and you want to discover whether or not it is infinitely divisible. Calculating the characteristic function, $\phi(u) = \sum_{k \geq 0} e^{iuk}(1 -$

$q)q^k = (1-q)/(1-qe^{iu})$, so $\log\phi(u) = \log(1-q) - \log(1-qe^{iu}) = -\sum_{j\geq 1} q^j/j + \sum_{j\geq 1}(q^j e^{iuj}/j) = \sum_{j\geq 1}((e^{iuj}-1)/j)\ q^j$.

Thus the geometric distribution has a Lévy representation in which ν has mass q^j at $j = 1, 2, \ldots$, so we have verified that the geometric distribution is infinitely divisible. The total mass a of ν is $a = q + q^2 + \cdots = q/(1-q)$; and this agrees with the recipe $a = \mathbb{E}X$. Since $\mathbb{P}(Y = j) = \nu(\{j\})/a = (1-q)q^{j-1}$ for $j = 1, 2, \ldots$, we have $Y \overset{\mathrm{d}}{=} 1 + X$. Thus $X^* = X + Y$ with X, Y independent and $Y \overset{\mathrm{d}}{=} X + 1$.

4b) Multiplying the Lévy measure ν by $t > 0$ yields the general case of the negative binomial distribution, $X \sim$ negative binomial(t, q). The case $t = 1$ is the geometric distribution. We still have $X^* = X + Y$ with X, Y independent, and $Y \sim$ geometric$(q) + 1$. Note that for integer t we can verify our calculation in another way, as in this case X is the sum of t independent geometric(q) variables X_i. By (12), we can size bias X by size biasing a single geometric term, which is the same as adding an independent Y with distribution, again, $1 +$ geometric(q).

Example 5. Y is exponential. X is exponential or Gamma

5a) Let X be exponentially distributed with $\mathbb{E}X = 1/\alpha$, i.e. $\mathbb{P}(X > t) = e^{-\alpha t}$ for $t > 0$. As we saw in section 5 for the case $\alpha = 1$, $X^* = X + Y$ with X, Y independent and $Y \overset{\mathrm{d}}{=} X$. The case with general $\alpha > 0$ is simply a spatial scaling of the mean one, "standard" case. The Lévy measures ν is simply $\mathbb{E}X = 1/\alpha$ times the common distribution of X and Y, with $\nu(dy) = e^{-\alpha y}\ dy$

5b) Multiplying the Lévy measure ν by $t > 0$ yields the general case of the Gamma distribution, $X \sim$ Gamma(α, t). The name comes from that fact that X has density $f(x) = (\alpha^t/\Gamma(t))\ x^{t-1}e^{-\alpha x}$ on $(0, \infty)$. The special case $t = 1$ is the exponential distribution, and more generally the case $t = n$ can be realized as $X = X_1 + \cdots + X_n$ where the X_i are iid, exponentially distributed with $\mathbb{E}X_i = 1/\alpha$. We have $X^* = X + Y$ with X, Y independent and Y is exponentially distributed with mean $(1/\alpha)$, so that $X^* \sim$ Gamma$(\alpha, t{+}1)$. The Lévy measure here is ν with $\nu(dy) = te^{-\alpha y}\ dy$; so the corresponding μ has $\mu(dy) = te^{-\alpha y}/y\ dy$. This form of μ is known as the Moran or Gamma subordinator; see e.g. [13]. As in example 4b), for integer t we can verify our calculation by noting that X is the sum of t independent exponential, mean $(1/\alpha)$ variables, and that by (12), when size biasing we will get the same Y added on to the sum as the Y which appears when size biasing any summand.

Example 6. Y **is ??,** X **is lognormal**

As mentioned in Section (15), we say that X is lognormal when $X = e^{\sigma Z+\mu}$ where Z is a standard normal variable. The proof that the lognormal is infinitely divisible, first given by Thorin [19], remains difficult; there is an excellent book by Bondesson [7] for further study. Consider even the standard case, $X = e^Z$, so that by equation (15), $X^* \overset{\mathrm{d}}{=} e^{Z+1} \overset{\mathrm{d}}{=} eX$. The result of Thorin that this X is infinitely divisible is thus equivalent, by Theorem 11.1, to the claim that there exists a distribution for $Y \geq 0$ such that with X and Y independent, $X + Y \overset{\mathrm{d}}{=} eX$. Also, by Theorem 11.1 with $a = \mathbb{E}X = \sqrt{e}$, the distribution of Y is exactly $1/\sqrt{e}$ times the Lévy measure ν. However, there does not seem to exist yet a simplified expression for this distribution!

Since the lognormal $X = e^Z$ satisfies $X^* = eX = X + (1-e)X$, it provides a simple illustration of our remarks in the paragraph following (6), that the relation $X^* = X + Y, Y \geq 0$ does not determine the distribution of Y without the further stipulation that X and Y be independent. Note also that in $X^* = X + (1-e)X$, the increment $(1-e)X$ is a monotone function of X, so this is an example of the coupling using the quantile transformation.

Example 7. Y **is uniform on an interval** (β, γ)**, with** $0 \leq \beta < \gamma < \infty$**.** By scaling space (dividing by γ) we can assume without loss of generality that $\gamma = 1$. This still allows two qualitatively distinct cases, depending on whether $\beta = 0$ or $\beta > 0$.

Example 7a. $\beta = 0$**: Dickman's function and its convolution powers**.

With $a = \mathbb{E}X \in (0, \infty)$, this example is specified by (26) with ν being a times the uniform distribution on (0,1), so that $\mu(dx) = a/x\ dx$ on $(0,1)$. The reader must take on faith that ν having a density, together with $\mu(\ (0,\infty)\ ) = \infty$ so that $\mathbb{P}(X = 0) = 0$, implies that the distribution of X has a density, call it g_a. Size biasing then gives an interesting differential-difference equation for this density: using (32), for $x > 0$,

$$g_a(x) = \frac{a}{x} \int_{y=0}^{1} g_a(x-y)\ dy = \frac{a}{x} \int_{x-1}^{x} g_a(z)\ dz. \qquad (34)$$

Multiplying out gives $x g_a(x) = a \int_{x-1}^{x} g_a(z)\ dz$, and taking the derivative with respect to x yields $x g_a'(x) + g_a(x) = a g_a(x) - a g_a(x-1)$, so that $g_a'(x) = (\ (a-1) g_a(x) - a g_a(x-1)\ )/x$, for $x > 0$.

For the case $a = 1$ this simplifies to $g_1'(x) = -g_1(x-1)/x$, which is the same differential-difference equation that is used to specify Dickman's function ρ, of central importance in number theory; see [18]. The function ρ is characterized by $\rho(x) = 1$ for $0 \leq x \leq 1$ and $\rho'(x) = -\rho(x-1)/x$ for $x > 0$, and $\rho(x) = 0$ for $x < 0$, with ρ continuous on $[0, \infty)$, and from the calculation that $\int_0^\infty \rho(x)\ dx = e^\gamma$, where γ is Euler's constant, it follows that $g_1(x) = e^{-\gamma}\rho(x)$. Dickman's function governs the distribution of the largest prime factor of a random integer in the following sense: for fixed $u > 0$, the proportion of integers from 1 to n whose largest prime factor is smaller than $n^{1/u}$ tends to $\rho(u)$ as $n \to \infty$. For example, $\rho(2)$ can be calculated from the differential equation simply by $\rho(2) = \rho(1) + \int_1^2 \rho'(x)\ dx = 1 + \int_1^2 -\rho(x-1)/x\ dx$ $= 1 + \int_1^2 -1/x\ dx = 1 - \log 2 \doteq 1 - .69314 = .30686$, and the claim is that $\rho(2)$ gives, for large n, the approximate proportion of integers from 1 to n all of whose prime factors are at most $\sqrt{n}$.

For general $t > 0$ the density g_t is a "convolution power of Dickman's function," see [12]. The size bias treatment of this first appeared in the 1996 version of [2], and was subsequently written up in [1].

Example 7b. $\beta > 0$: Buchstab's function, integers free of small prime factors.

For these examples Y is uniform on $(\beta, 1)$ for $\beta \in (0,1)$, with density $1/(1-\beta)$ on $(\beta, 1)$. Therefore ν is a multiple of uniform distribution on $(\beta, 1)$, with density t on $(\beta, 1)$ for some constant $t > 0$ — we have $a := \mathbb{E}X = t(1-\beta)$ — but t rather than a is the convenient parameter. From $\nu(dx) = t\ dx$ on $(\beta, 1)$ we get $\mu(dx) = t/x\ dx$ on $(\beta, 1)$, so that the total mass of μ is $\lambda = \int_{(\beta,1)} t/x\ dx = t\log(1/\beta)$. Since $\lambda < \infty$, X is compound Poisson with $\mathbb{P}(X = 0) = e^{-\lambda} = \beta^t$.

For the case $t = 1$, the distribution of the random variable X is related to another important function in number theory, Buchstab's function ω; again see [18]. The relation involves a "defective density" — here $t = 1$ and $\mathbb{P}(X = 0) = \beta > 0$ so X does not have a proper density. Size biasing yields a relation similar to (32), which leads to a differential-difference equation, which in turn establishes the relation between the defective density and Buchstab's function; see [3]. The net result is that for $\beta < a < b < 1$, $\mathbb{P}(a < X < b) = \int_a^b \omega(x/\beta)\ dx$. Buchstab's function ω is characterized by the properties that it is continuous on $(1, \infty)$, $\omega(u) = 1/u$ for $u \in [1, 2]$, and $(u\omega(u))' = \omega(u-1)$ for $u > 2$. It governs the distribution of the smallest prime factor of a random integer in the sense that for $u > 1$, the proportion of integers form 1 to n whose smallest prime factor is at least $n^{1/u}$ is asymptotic to $u\omega(u)/\log n$.

References

[1] Arratia, R. (1998) On the central role of the scale invariant Poisson processes on $(0,\infty)$. In D. Aldous and J. Propp editors, Microsurveys in Discrete Probability, pages 21-41, DIMACS Series in Discrete Math. and Theoret. Comput. Sci., Amer. Math. Soc., Providence RI.

[2] Arratia, R., Barbour, A. D., and Tavaré, S. (1997) Logarithmic combinatorial structures. Monograph, 187 pages, in preparation.

[3] Arratia, R., and Stark, D. (1998) A total variation distance invariance principle for primes, permutations and Poisson-Dirichlet. Preprint.

[4] Baldi, P. Rinott, Y. (1989). On normal approximations of distributions in terms of dependency graphs, Annals of Probability **17** , 1646-1650.

[5] Baldi, P. Rinott, Y. and Stein, C. (1989). A normal approximations for the number of local maxima of a random function on a graph. In Probability, Statistics and Mathematics, Papers in Honor of Samuel Karlin. T. W. Anderson, K.B. Athreya and D. L. Iglehart eds., Academic Press , 59-81.

[6] Barbour, Holst, and Janson (1992). Poisson Approximation. Oxford Science Publications.

[7] Bondesson, L. (1992) Generalized Gamma Convolutions and Related Classes of Distributions and Densities. Lecture Notes in Statistics, vol. 76. Springer.

[8] Cochran, W. (1977) Sampling Techniques John Wiley & Sons, New York.

[9] Feller, W. (1966) An Introduction to Probability and its Applications, volume II. Wiley.

[10] Goldstein, L. and Rinott, Y. (1996) Multivariate normal approximations by Stein's method and size bias couplings. Journal of Applied Probability **33**, 1-17.

[11] van Harn, K. and Steutel, F. W. (1995) Infinite divisibility and the waiting-time paradox. Comm. Statist. Stochastic Models 11 (1995), no. 3, 527–540.

[12] Hensley, D. (1986). The convolution powers of the Dickman function. J. London Math. Soc. (2) **33**, 395-406.

[13] Kingman, J.F.C. (1993) Poisson Processes. Oxford Science Publications

[14] Lyons, R., Pemantle, R., and Peres, Y. (1995) Conceptual proof of $L \log L$ criteria for mean behavior of branching processes. Ann. Probab. **23**, 1125-1138.

[15] Perman, M., Pitman, J., and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. Probab. Th. Rel. Fields **92**, 21-39.

[16] Slud, E. (1993) The moment problem for polynomial forms in normal random variables. Ann. Probab. **21**, 2200-2214.

[17] Steutel, W. F. (1973). Some recent results in infinite divisibility. Stoch. Pr. Appl. **1**, 125-143.

[18] Tenenbaum, G. (1995) Introduction to analytic and probabilistic number theory. Cambridge studies in advanced mathematics, **46**. Cambridge University Press.

[19] Thorin, O. (1977) On the infinite divisibility of the lognormal distribution. Scand. Actuarial J., 121-148.